(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－3321

(43)公開日　平成11年(1999)１月６日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　15/16	３７０	Ｇ０６Ｆ　15/16	３７０Ｎ
		9/46	３６０Ｂ
9/46	３６０	15/16	４２０Ｊ

審査請求　未請求　請求項の数３　ＯＬ　（全７頁）

(21)出願番号　特願平9－155374

(22)出願日　平成９年(1997)６月12日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

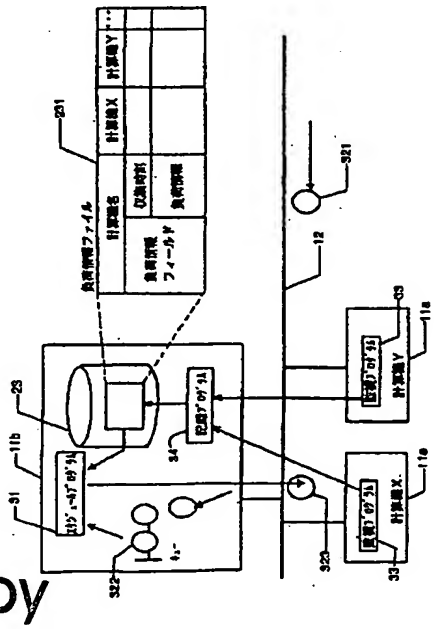
(72)発明者　平塚　正史
神奈川県横浜市戸塚区戸塚町5030番地　株式会社日立製作所ソフトウェア開発本部内

(74)代理人　弁理士　武　顕次郎

(54)【発明の名称】　並列計算機システム

(57)【要約】
【課題】　各ジョブ実行用計算機の負荷に応じてジョブを割り当てる並列計算機システムに関し、負荷を高めることなく負荷情報の収集を行う。
【解決手段】　各々のジョブ実行用計算機１１ａで、他のどのジョブよりも低い優先順位でそのジョブ実行用計算機１１ａの負荷情報を収集してスケジューラ計算機１１ｂに送信する監視プログラム３３を実行させる。また、スケジューラ計算機１１ｂで、監視プログラム３３から送信された最新の負荷情報をその収集時刻とともにジョブ実行用計算機１１ａごとに記録する記録プログラム３４と、新たなジョブが与えられたとき、記録プログラム３４により記録された負荷情報及びその収集時刻に基づいてこの新たなジョブを割り当てるジョブ実行用計算機１１ａを選択するスケジュールプログラム３１とを実行させる。

【図２】



Best Available Copy

1

【特許請求の範囲】
【請求項１】　割り当てられたジョブを実行する複数のジョブ実行用計算機と、
各々の前記ジョブ実行用計算機にネットワーク経由で接続され、新たなジョブをいずれかの前記ジョブ実行用計算機または自スケジューラ計算機に割り当てるスケジューラ計算機とを具備し、
前記スケジューラ計算機が、各々の前記ジョブ実行用計算機の負荷を表す負荷情報に基づき、新たなジョブを割り当てる計算機を少なくとも１つ選択する並列計算機システムにおいて、
各々の前記ジョブ実行用計算機は、
他のどのジョブよりも低い優先順位で自ジョブ実行用計算機の前記負荷情報を収集して前記スケジューラ計算機に送信する負荷情報収集送信手段を備え、
前記スケジューラ計算機は、
前記負荷情報収集送信手段から送信された最新の前記負荷情報をその収集時刻とともに前記ジョブ実行用計算機ごとに記録する負荷情報記録手段と、
新たなジョブが与えられたとき、前記負荷情報記録手段により記録された前記負荷情報及びその収集時刻に基づいて前記新たなジョブを割り当てる計算機を選択する計算機選択手段とを備えることを特徴とする並列計算機システム。
【請求項２】　前記計算機選択手段は、前記負荷情報記録手段により記録された前記負荷情報及びその収集時刻に基づき、前記収集時刻から現在までに経過した時間を前記ジョブ実行用計算機ごとに求め、
この時間が所定の収集時間間隔を超過しているすべての前記ジョブ実行用計算機を、新たなジョブの割り当て対象から除外することを特徴とする請求項１記載の並列計算機システム。
【請求項３】　前記計算機選択手段は、新たなジョブを割り当てる計算機として、前記時間が前記収集時間間隔を超過していない前記ジョブ実行用計算機のうち、対応する前記負荷情報で表される負荷が最小のジョブ実行用計算機を選択することを特徴とする請求項２記載の並列計算機システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明はジョブを複数の計算機に分散実行させる並列計算機システムに係り、特に、各計算機の負荷情報に応じてジョブの割当てを行う並列計算機システムに関する。
【０００２】
【従来の技術】従来より、複数のジョブ実行用計算機及びこれらのジョブ実行用計算機にジョブを割り当てるスケジューラ計算機をネットワーク経由で相互に接続し、新たに投入されたジョブをスケジューラ計算機が各々のジョブ実行用計算機に分散実行させる並列計算機システムが知られている。

2

こうした並列計算機システムでは、ジョブが投入されたとき、負荷が最も小さいと考えられるジョブ実行用計算機をスケジューラ計算機が選択し、このジョブ実行用計算機に対してジョブの実行を割り当てる、という負荷分散処理を行っている。そして、この負荷分散処理を実現するには、スケジューラ計算機が、各々のジョブ実行用計算機からそのジョブ実行用計算機の負荷を表す負荷情報を収集することが必要である。このため、スケジューラ計算機は、一定時間ごとに、あるいは新たなジョブが投入されたときに、この負荷情報の収集を行っている。
【０００３】
【発明が解決しようとする課題】上述した負荷情報の収集を一定時間ごとに行う方法は、ジョブ実行用計算機の負荷が既にかなり高くなっているときでも負荷情報の収集が行われるため、この負荷情報の収集によってジョブ実行用計算機やネットワークの負荷をさらに高めてしまい、これによって、並列計算機システムの性能低下を招くなどの悪影響を及ぼすことがあるという問題点があった。
【０００４】一方、新たなジョブが投入されたときに負荷情報の収集を行う方法は、新たなジョブが投入された時点で、あらゆるジョブ実行用計算機に対して同時に負荷情報の収集が行われるため、ネットワークの負荷を急激に高め、また、それらの収集された負荷情報を解析するスケジューラ計算機に対して短時間に負荷が集中してしまい、これによって、並列計算機システムのスループットが低下することがあるという問題点があった。
【０００５】したがって本発明の目的は、上記方法を用いる従来技術の問題点を解決して、新たなジョブの割当て対象とするジョブ実行用計算機の選択に必要な負荷情報の収集に際し、ジョブ実行用計算機やネットワークの負荷が急激に高められることのない並列計算機システムを提供することにある。
【０００６】
【課題を解決するための手段】上記の目的を達成するため、本発明の並列計算機システムは、割り当てられたジョブを実行する複数のジョブ実行用計算機と、各々の前記ジョブ実行用計算機にネットワーク経由で接続され、新たなジョブをいずれかの前記ジョブ実行用計算機または自スケジューラ計算機に割り当てるスケジューラ計算機とを具備し、前記スケジューラ計算機が、各々の前記ジョブ実行用計算機の負荷を表す負荷情報に基づき、新たなジョブを割り当てる計算機を少なくとも１つ選択する並列計算機システムにおいて、各々の前記ジョブ実行用計算機は、他のどのジョブよりも低い優先順位で自ジョブ実行用計算機の前記負荷情報を収集して前記スケジューラ計算機に送信する負荷情報収集送信手段を備え、前記スケジューラ計算機は、前記負荷情報収集送信手段から送信された最新の前記負荷情報をその収集時刻とと

もに前記ジョブ実行用計算機ごとに記録する負荷情報記録手段と、新たなジョブが与えられたとき、前記負荷情報記録手段により記録された前記負荷情報及びその収集時刻に基づいて前記新たなジョブを割り当てる計算機を選択する計算機選択手段とを備えるものである。

【０００７】

【発明の実施の形態】以下、本発明の並列計算機システムの実施の形態を図面を用いて詳細に説明する。

【０００８】図１は本発明の並列計算機システムの一実施形態の構成を示すブロック図、図２は図１のシステムにおける負荷情報の収集について説明する概念図である。両図中、１１ａは割り当てられたジョブを実行する複数のジョブ実行用計算機、１１ｂは各々のジョブ実行用計算機１１ａにネットワーク１２経由で接続され、新たなジョブをジョブ実行用計算機１１ａのいずれかに割り当てるスケジューラ計算機、２１はスケジューラ計算機１１ｂで実行されるプログラムに応じた処理を行うＣＰＵ、２２はスケジューラ計算機１１ｂで処理中のデータなどを一時的に格納するメモリ、２３はスケジューラ計算機１１ｂでプログラムやデータを格納する外部記憶装置、２３１は外部記憶装置２３に設けられ、後述する記録プログラムにより各々のジョブ実行用計算機１１ａの負荷を表す負荷情報及びその収集時刻が記録される負荷情報ファイル、２４はこれらＣＰＵ２１，メモリ２２，外部記憶装置２３の間でプログラムやデータをやりとりするためのバス、３１はスケジューラ計算機１１ｂのＣＰＵ２１で常に実行され、上述した新たなジョブをジョブ実行用計算機１１ａのいずれかに割り当てる具体的な処理を行う計算機選択手段であるスケジュールプログラム、３２１は新たに投入されたジョブ、３２２は割当て待ちキューに登録された実行割当て待ちジョブ、３２３は割り当てられた特定のジョブ実行用計算機１１ａへ送信中の実行待ちジョブ、３３は各々のジョブ実行用計算機１１ａで実行される負荷情報収集送信手段である監視プログラム、３４はスケジューラ計算機１１ｂのＣＰＵ２１で常に実行され、上述の監視プログラム３３から送信された最新の負荷情報をその収集時刻とともにジョブ実行用計算機１１ａごとに負荷情報ファイル２３１に記録する負荷情報記録手段である記録プログラムである。なお、図１中にはジョブ実行用計算機１１ａを４台のみ接続した構成を示してあるが、このジョブ実行用計算機１１ａの接続台数は、必要に応じて所望の台数とすることができる。

【０００９】図２において、各々のジョブ実行用計算機１１ａで実行されている監視プログラム３３は、他のどのジョブよりも低い優先順位で、そのジョブ実行用計算機の負荷情報、例えばＣＰＵ使用率やメモリ使用率などを収集して前記スケジューラ計算機に送信する。この監視プログラム３３による負荷情報の収集及び送信は、一応の目安である後述する所定の収集時間間隔で繰り返し

行われる。すなわち、あるジョブ実行用計算機１１ａに割り当てられたジョブによるＣＰＵ使用率が比較的小さかった場合、優先順位の低い監視プログラム３３に実行制御が渡される頻度が大きくなるため、監視プログラム３３による負荷情報の収集及びそのスケジューラ計算機１１ｂへの送信は、概ね収集時間間隔ごとに行われる。これに対して、上記ジョブによるＣＰＵ使用率がある水準を超えて大きくなった場合、優先順位の低い監視プログラム３３に実行制御が渡される頻度は小さくなるため、監視プログラム３３による負荷情報の収集及びそのスケジューラ計算機１１ｂへの送信は、収集時間間隔より長い間隔で行われる。また、スケジューラ計算機１１ｂで実行されている記録プログラム３４は、各々のジョブ実行用計算機１１ａからネットワーク１２経由で送信されてきた上述の負荷情報を、その収集時刻とともにジョブ実行用計算機１１ａごとに負荷情報ファイル２３１中に記録する。すなわち、送信された負荷情報の発信元であるジョブ実行用計算機１１ａが“計算機Ｘ”であれば、負荷情報ファイル２３１中の“計算機Ｘ”に対応する負荷情報フィールド内に、送信された負荷情報をその収集時刻、例えばこの負荷情報をネットワーク１２から受信したときの時刻とともに記録する。

【００１０】図２において、スケジューラ計算機１１ｂで実行されているスケジュールプログラム３１は、新たに投入されたジョブ３２１が割当て待ちキューに登録されて実行割当て待ちジョブ３２２となったとき、記録プログラム３４により記録された負荷情報ファイル２３１を参照し、各々のジョブ実行用計算機１１ａごとに、記録された最新の収集時刻から現在時刻すなわち負荷情報ファイル２３１を参照した時刻までに経過した時間を求める。そして、この時間が所定の収集時間間隔を超過しているジョブ実行用計算機１１ａは、すべて実行割当て待ちジョブ３２２の割り当て対象から除外する。続いて、この時間が所定の収集時間間隔を超過していないジョブ実行用計算機１１ａについて負荷情報ファイル２３１を参照し、対応する負荷情報で表される負荷が最小のジョブ実行用計算機１１ａを、新たな実行割当て待ちジョブ３２２を割り当てるべきジョブ実行用計算機１１ａとして選択する。最後に、この実行割当て待ちジョブ３２２を実行待ちジョブ３２３として、ネットワーク１２経由で選択された特定のジョブ実行用計算機１１ａへ送信する。

【００１１】図３は、図１のシステムにおけるジョブの割当て対象の選択について説明する概念図である。同図中、横軸は時刻の経過を表し、ｔは現在時刻、ｉは一応の目安である所定の収集時間間隔である。ここで、収集時間間隔ｉの具体的な時間の値は、並列計算機システムの運用方針などに基づいて決めた適当な定数値としてある。また、ｄは、負荷情報ファイル２３１に記録された最新の収集時刻から現在時刻ｔまでに経過した時間であ

5

る。

【００１２】一般に、ジョブ実行用計算機１１ａに割り当てられたジョブによる負荷は、ジョブの同時実行数やジョブの処理状態などに応じて常に変化する。そして、上述したように、監視プログラム３３が負荷情報の収集及び送信を行う間隔は、ジョブ実行用計算機１１ａに割り当てられたジョブによるＣＰＵ使用率などの負荷に応じて常に変化する。すなわち、この負荷が比較的小さいとき、負荷情報の収集及び送信の間隔は収集時間間隔ｉに概ね一致するが、負荷がある水準を超えて大きくなると、上記間隔は収集時間間隔ｉより長くなる。このように、負荷情報の収集及び送信の間隔は常に変化しているが、少なくともこの間隔が収集時間間隔ｉより長い場合、該当するジョブ実行用計算機１１ａに割り当てられたジョブによる負荷がある水準を超えて大きくなっているものとみなされる。すなわち、図３に示すように、負荷情報ファイル２３１に記録された最新の収集時刻から現在時刻ｔまでに経過した時間ｄと収集時間間隔ｉとの大小関係に対応して、３種類のケース①②③が生ずるが、このうち、最新の収集時刻から現在時刻ｔまでに経過した時間ｄが収集時間間隔ｉより長いケース③は、該当するジョブ実行用計算機１１ａに割り当てられたジョブによる負荷がある水準を超えて大きくなっているものとみなされる。したがって、このケース③に該当するジョブ実行用計算機１１ａは、すべて新たなジョブの割り当て対象から除外する。次に、残りのケース①②については、該当するジョブ実行用計算機１１ａに割り当てられたジョブによる負荷がある水準より小さいことしかわからないので、再び負荷情報ファイル２３１を参照する。そして、記録された負荷情報で表される負荷が最小となっているジョブ実行用計算機１１ａを、新たなジョブを割り当てるジョブ実行用計算機１１ａとして選択する。

【００１３】以上のように、本実施形態の並列計算機システムにおいて、各々のジョブ実行用計算機１１ａで実行されている監視プログラム３３は、他のどのジョブよりも低い優先順位でそのジョブ実行用計算機１１ａの負荷を表す負荷情報を収集してスケジューラ計算機１１ｂに送信している。このため、通常の優先順位のジョブの実行に伴ってジョブ実行用計算機１１ａの負荷がかなり高くなったときは、より低い優先順位の監視プログラム３３の実行はほとんど停止し、負荷情報の収集及びスケジューラ計算機１１ｂへの送信が事実上行われなくなるので、既にかなり負荷が高くなっているジョブ実行用計算機１１ａやネットワーク１２の負荷が監視プログラム３３によってさらに高められることがない。また、スケジューラ計算機１１ｂで実行されているスケジュールプログラム３１は、新たなジョブが与えられたとき、既に負荷情報ファイル２３１に記録されている負荷情報のみを参照して、このジョブを割り当てるジョブ実行用計算機１１ａを選択し、各々のジョブ実行用計算機１１ａから改めて負荷情報を収集しないため、新たなジョブの登録に際してジョブ実行用計算機１１ａやネットワーク１２の負荷が急激に高められることがない。したがって、新たなジョブの割当て対象とするジョブ実行用計算機１１ａの選択に必要な負荷情報の収集に際し、ジョブ実行用計算機１１ａやネットワークの負荷１２が急激に高められることがなくなり、並列計算機システムの性能低下などの悪影響の発生を防止することができる。

6

【００１４】さらに、新たなジョブの実行を割り当てるジョブ実行用計算機１１ａの選択に際し、各ジョブ実行用計算機１１ａに対応する最新の収集時刻から現在時刻ｔまでに経過した時間ｄと所定の収集時間間隔ｉとの比較によって割当て対象を絞り込んだ後、負荷情報で表される負荷が最小となっているジョブ実行用計算機１１ａを選択するので、前回収集された負荷情報による負荷は比較的小さかったが、現時点では負荷がある水準より大きいために時間ｄが収集時間間隔ｉを超過するジョブ実行用計算機１１ａを、割当て対象から除外することができる。

【００１５】なお、上述した実施形態では、スケジューラ計算機１１ｂが、新たなジョブをジョブ実行用計算機１１ａのいずれかに対してのみ割り当てる場合について説明を行ったが、割り当て対象にスケジューラ計算機１１ｂ自身を含めるようにしてもよい。この場合には、スケジューラ計算機１１ｂのすべてが上述したケース③に該当したとき、スケジューラ計算機１１ｂ自身を新たなジョブの割り当て対象として選択するのが適切である。

【００１６】

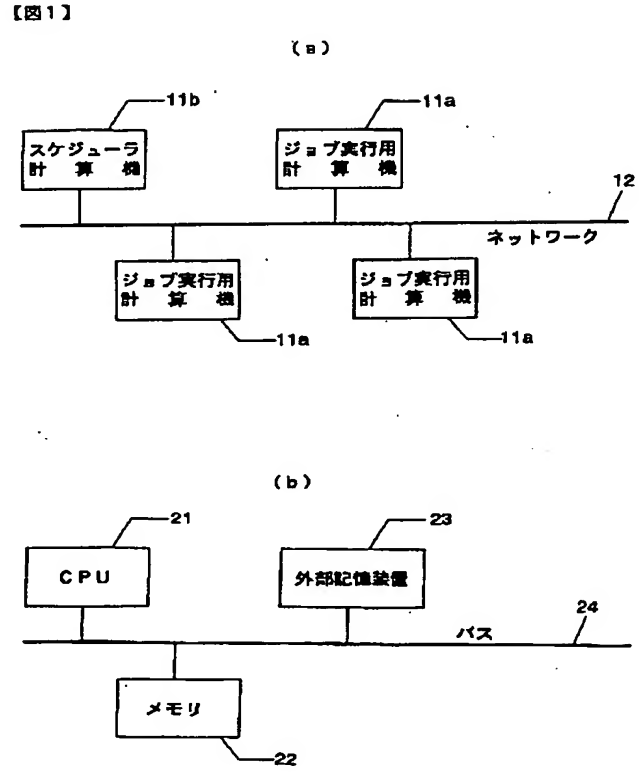
【発明の効果】以上詳しく説明したように、本発明の並列計算機システムによれば、通常の優先順位のジョブの実行に伴ってジョブ実行用計算機の負荷がかなり高くなったときは、より低い優先順位の負荷情報収集送信手段による負荷情報の収集及びスケジューラ計算機への送信が事実上行われなくなるので、既にかなり負荷が高くなっているジョブ実行用計算機やネットワークの負荷が負荷情報収集送信手段によってさらに高められることがない。また、計算機選択手段は、新たなジョブが与えられたとき、負荷情報記録手段によって既に負荷情報ファイルに記録されている負荷情報のみを参照して、このジョブを割り当てるジョブ実行用計算機を選択し、各々のジョブ実行用計算機から改めて負荷情報を収集しないため、新たなジョブの登録に際してジョブ実行用計算機やネットワークの負荷が急激に高められることがない。したがって、新たなジョブの割当て対象とするジョブ実行用計算機の選択に必要な負荷情報の収集に際し、ジョブ実行用計算機やネットワークの負荷が急激に高められることがなくなり、並列計算機システムの性能低下などの悪影響の発生を防止することができる。

【図面の簡単な説明】

7　　8

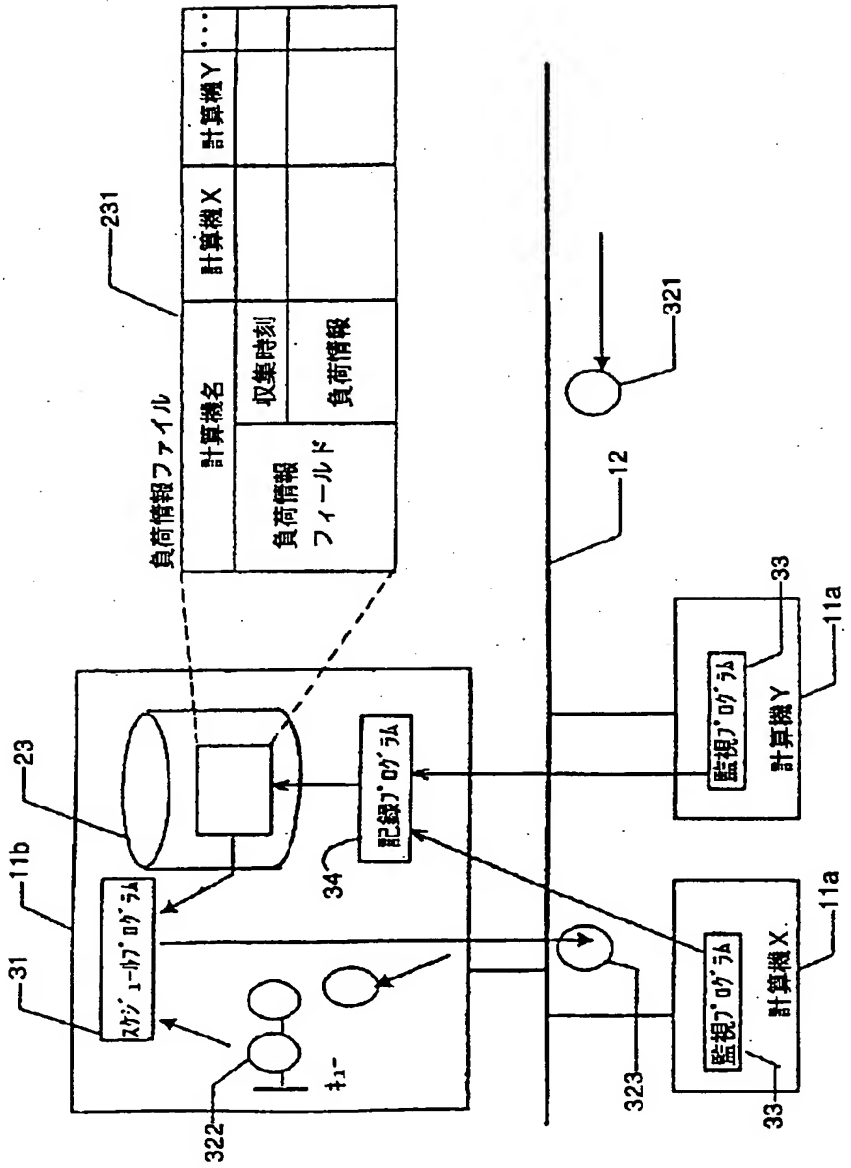
【図１】本発明の並列計算機システムの一実施形態の構成を示すブロック図である。
【図２】図１のシステムにおける負荷情報の収集について説明する概念図である。
【図３】図１のシステムにおけるジョブの割当て対象の選択について説明する概念図である。
【符号の説明】
１１ａ　ジョブ実行用計算機
１１ｂ　スケジューラ計算機
＊１２　ネットワーク
２１　ＣＰＵ
２３　外部記憶装置
２３１　負荷情報ファイル
３１　スケジュールプログラム
３２２　実行割当て待ちジョブ
３２３　実行待ちジョブ
３３　監視プログラム

【図１】



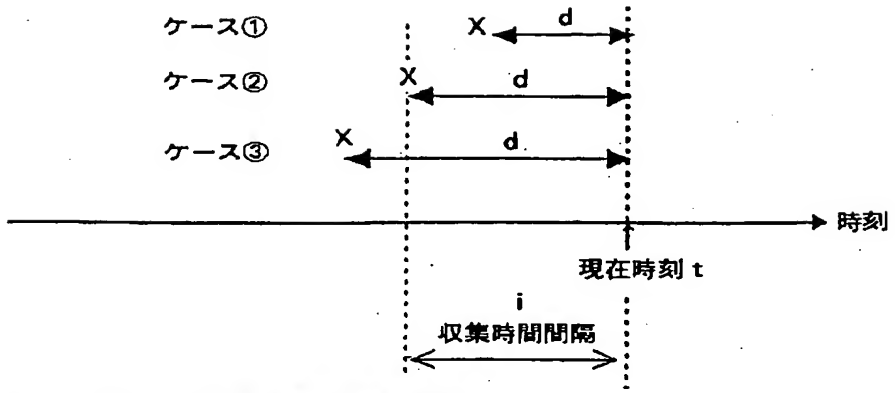
【図２】

【図２】


231
負荷情報ファイル
計算機名
収集時刻
負荷情報
負荷情報フィールド
計算機X
計算機Y
321
12
33
11a
監視ﾌﾟﾛｸﾞﾗﾑ
計算機Y
23
11b
34
記録ﾌﾟﾛｸﾞﾗﾑ
31
ｽｹｼﾞｭｰﾙﾌﾟﾛｸﾞﾗﾑ
ｷｭｰ
322
323
監視ﾌﾟﾛｸﾞﾗﾑ
計算機X
33
11a

【図３】

【図３】



X　：各ケースの負荷情報の収集時刻
⟷　：ユーザから指定されている収集時間間隔
◀⟷▶　：収集時刻から現在時刻までの間に経過した時間

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☑ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.